Published online 6 December 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D297–D303***
doi:10.1093/nar/gkt1208

MMDB and VAST+: tracking structural similarities between macromolecular complexes

Thomas Madej*, Christopher J. Lanczycki, Dachuan Zhang, Paul A. Thiessen, Renata C. Geer, Aron Marchler-Bauer and Stephen H. Bryant

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38 A, Room 8N805, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received November 3, 2013; Accepted November 4, 2013

ABSTRACT

The computational detection of similarities between protein 3D structures has become an indispensable tool for the detection of homologous relationships, the classification of protein families and functional inference. Consequently, numerous algorithms have been developed that facilitate structure comparison, including rapid searches against a steadily growing collection of protein structures. To this end, NCBI's Molecular Modeling Database (MMDB), which is based on the Protein Data Bank (PDB), maintains a comprehensive and up-to-date archive of protein structure similarities computed with the Vector Alignment Search Tool (VAST). These similarities have been recorded on the level of single proteins and protein domains, comprising in excess of 1.5 billion pairwise alignments. Here we present VAST+, an extension to the existing VAST service, which summarizes and presents structural similarity on the level of biological assemblies or macromolecular complexes. VAST+ simplifies structure neighboring results and shows, for macromolecular complexes tracked in MMDB, lists of similar complexes ranked by the extent of similarity. VAST+ replaces the previous VAST service as the default presentation of structure neighboring data in NCBI's Entrez query and retrieval system. MMDB and VAST+ can be accessed via http://www.ncbi.nlm.nih.gov/Structure.

INTRODUCTION

NCBI has maintained the Molecular Modeling Database (MMDB) (1) since 1996, as a collection of publicly accessible experimentally determined macromolecular structures that have been deposited with the Protein Data Bank (PDB) (2). MMDB serves a variety of functions. It facilitates searching for macromolecular structure data in NCBI's Entrez query and retrieval system (3); links and associates macromolecular structure data with a variety of other resources such as gene, sequence, sequence variation, chemistry and literature databases; provides sequence data for NCBI's BLAST (4) services; and supports other NCBI resources such as the Conserved Domain Database (CDD) (5) and IBIS (6). MMDB mirrors the content of PDB and currently contains 94 688 macromolecular structures. More than 97% of these structures have at least one protein or polypeptide component. Recently, MMDB has updated the default presentation of macromolecular structure data so that the biologically relevant macromolecular complexes (termed 'biological units' or 'biological assemblies'), as defined by the structures' authors or computed by the PDB (7), are shown by default, and so that interactions between macromolecules or macromolecules and smaller chemical ligands are emphasized (1). This presentation makes it easier to identify pairs of molecules that come into direct contact with each other in a macromolecular complex—something that can be difficult to determine, for example, when visualizing large biological assemblies using 3D graphics software.

Another feature of MMDB is the association of 3D structures, through sequence similarity searches, with protein sequences that do not yet have solved structures, facilitating inference of protein function. A BLAST search of all sequences in the Entrez Protein database, against the subset of protein sequences from experimentally determined structures, maps a large fraction of the publicly available proteins to three-dimensional structure information in MMDB. For example, of the 35 138 human protein sequences tracked in NCBI's BioProject 178 030, a genomic sequence data set from a human hydatidiform mole cell line, 76% appear similar to known 3D structures using standard protein-BLAST. An even larger fraction can be mapped to 3D structure using approaches with higher sensitivity, such as the identification of conserved domain signatures. Thus,

*To whom correspondence should be addressed. Tel: +1 301 435 5998; Fax: +1 301 435 7793; Email: madej@ncbi.nlm.nih.gov

macromolecular structure data as provided by MMDB can be used to postulate homology-inferred function for a large number of functionally uncharacterized protein sequences and genes.

What may be less well-known and used are structure neighboring data and a structure neighboring service available as part of MMDB, which identify similarly-shaped structures based on geometric criteria, regardless of the extent of sequence similarity. The resulting 3D structure alignments are helpful in understanding the functional consequences of sequence variation, as well as in discovering distant homologous relationships and subtle functional similarities. The Vector Alignment Search Tool (VAST) (8) algorithm, which computes these similarities, was developed around 1995 and has been applied ever since to compute and maintain comprehensive and up-to-date lists of statistically significant similarities between known protein 3D structures. Pre-computed similarities and alignments derived from structure superposition are available for all protein structures that have been included in MMDB and are suitable to be processed by VAST. An interactive search tool, VAST-Search, facilitates structure similarity searches for protein structure queries that are not (yet) part of MMDB's collection, enabling a user to enter 3D coordinate data for comparison against all publicly available structures.

In the past 25 years, a variety of methods have been developed to computationally characterize or measure structural similarities between macromolecules, resulting in an even larger variety of published methods, too numerous to be listed here (9). The Protein Data Bank, for example, reports structure neighbors from a subset of representatives computed with the jFATCAT algorithm (10) and points to a handful of external resources that provide structural classifications and structure comparisons: SCOP (11), CATH (12), VAST (8), FATCAT (13), DALI (14) and Superfamily (15). SCOP is a hierarchical classification of domain structures that has been maintained by manual intervention and does not rely on computationally determined 3D structure similarity. CATH classifies domain structures hierarchically as well, but makes systematic use of the SSAP (16) algorithm to compute similarities on a 3D level. FATCAT, a more recent development, uses dynamic programming to string together locally aligned pairs of structural fragments while allowing for a number of twists around pivot points, decomposing the match between two structures into a series of segment pairs that can be superimposed as rigid bodies. The jFATCAT implementation used to pre-compute data for the PDB site falls back to reporting a single-segment rigid body superimposition, though. DALI was one of the first structure comparison methods that relied solely on geometric criteria and has been available since the mid 1990s. Another method that had been incorporated in the Protein Data Bank, CE (17), computes rigid body superimpositions for alignments found via combinatorial extension of aligned fragment pairs, as opposed to dynamic programming or Monte Carlo optimizations.

Most, if not all of these computational resources and the associated data have been maintained and available/accessible since their inception, although updates of data sets such as pre-computed structure alignments may not have happened frequently, as most structure comparison methods are computationally intensive. The DALI database of pre-computed structure alignments, for example, currently reports a most recent update in March 2011. Also, most of the pre-computed structure neighboring data sets and search databases available for live neighboring have been reduced in size to contain representative structures only. Pre-computed structure neighbors as found on the PDB Web site, for example, have been obtained for representative structures from clusters formed at a threshold of 40% sequence identity, meaning that a structural alignment between an arbitrary pair of similar or related protein structures may not be readily available, which somewhat limits the practical applicability of the data and search implementations.

The VAST search database and database of pre-computed structure alignments have been maintained as complete and redundant collections since their launch, with automated updates occurring on a weekly basis. This was made possible by implementing a fast heuristic that uses a model for the statistical significance of initial alignments of secondary structure vectors (which can be computed quickly), so that the database searches can avoid costly alignment refinements for the large majority of insignificant and uninteresting similarities. The drawbacks are that a heuristic will miss some potentially interesting similarities. The VAST algorithm will not, for example, report similarities between structures deemed to have <3 secondary structure elements. Searches for structural similarity can and should be complemented with searches for sequence similarity, as flexibility of molecular structure and limitations of the structure comparison method may preclude the detection of matches between structures of homologous polypeptides. In general, though, structure comparison methods will pick up many subtle similarities that evade detection by sequence comparison strategies, and there is no natural cutoff point for a ranked list of similar structures, unlike in the sequence comparison scenario, where matches to non-homologous gene products are considered accidental and uninformative, for the most part.

Results computed by the VAST algorithm have been compared against other approaches a number of times (17–19). Although there are subtle differences in retrieval sensitivity and alignment accuracy (20), it appears fair to state that the large majority of extensive structural similarities, which are indicative of common evolutionary descent and could be used to infer functional similarities, are reported by VAST (and by most if not all of the alternative approaches to detect common substructures).

As structure similarity search strategies have been developed to also detect distant relationships that might not be evident from sequence analysis, most if not all of the current approaches have been implemented so that they use a single protein molecule or rather a single domain as the unit of comparison. This has been true for VAST, in particular. However, the Protein Data Bank is continuing to accumulate structures of larger macromolecular complexes and has started to provide data on what constitutes

functionally or biologically relevant macromolecular complexes or biological assemblies (1). Such assemblies range from simple homo-oligomers to intricate arrangements of many different components, revealing details on specific molecular interactions and on how these might constrain sequence variation. A small number of approaches have been published in the past few years that examine structural similarity of macromolecular complexes (21,22). Here we present a simple strategy that builds on the existing database of pairwise structure alignments computed by VAST and supports the first (to our knowledge) comprehensive and regularly updated collection of macromolecular complex similarities.

VAST+ AS AN EXTENSION TO EXISTING PROTEIN STRUCTURE COMPARISON

As information characterizing biological assemblies in macromolecular structure data has become available, it seemed that the biological assembly would be a convenient and informative unit of comparison between individual entries in the structure database. If the goal is to list structures most similar to any particular query, one would have to consider that the query itself may contain a macromolecular complex with a given stoichiometry, and that matching complexes with matching stoichiometry might be more informative 'structure neighbors' than, for example, the structures that happen to contain molecules with the strongest local similarity to the query, irrespective of the context.

VAST+ builds on the existing VAST database to generate such a report of structure neighbors. Its goal is to find the largest set of pairs of matching macromolecules between two biological assemblies and to characterize that match and compute instructions for a global superimposition that can be used to visualize the structural similarity. For each pair of structures in MMDB, VAST+ examines pre-computed structure alignments stored in the VAST database that were computed for the full-length protein molecule components of the default biological assemblies. If such pairwise alignments are found, the alignments between individual protein components of the biological assemblies are compared with each other for compatibility, and compatible/matching alignments are clustered into sets of alignments that together constitute a biological assembly match. Pairwise alignments are compatible (i) if they do not share the same macromolecules, i.e. a protein molecule from one assembly cannot be aligned to two molecules from the other assembly at the same time and (ii) if they generate similar instructions (spatial transformation matrices) for the superpositions of coordinate sets. A simple distance metric can be used to compare transformation matrices and it lends itself to cluster alignment sets efficiently.

Each set of compatible pairwise alignments can be characterized by (i) the number of pairwise matches, i.e. the total number of pairs of protein molecules from the query and subject biological assemblies, that are simultaneously aligned with each other; (ii) the RMSD of the superposition obtained from considering all alignments in the set; (iii) the total length of all pairwise alignments, i.e. the total number of amino acids that are aligned in 3D space; and (iv) percentage of identical residues in the alignments. For each pairwise comparison of two biological assemblies, only the match with the highest number of aligned molecules and the highest number of aligned residues is recorded and reported.

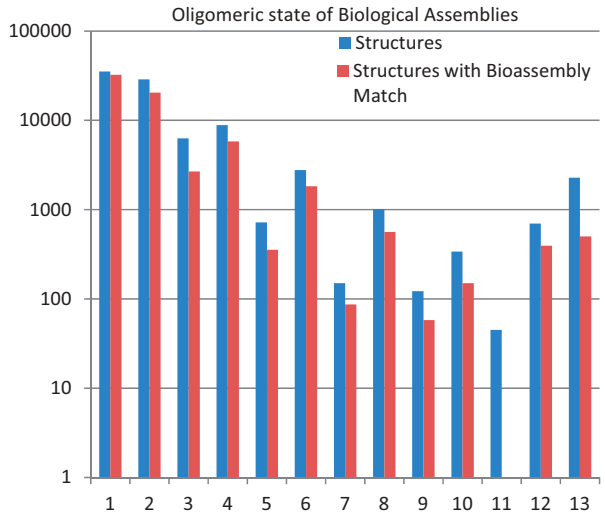

Figure 1. This histogram displays the number of structures in MMDB (blue), categorized by the size of the biological assembly. Monomers, dimers and higher oligomers up to dodecamers are plotted as separate categories, the 13th category summarizes tridecamers and all higher oligomers. The *y*-axis is scaled logarithmically. Red columns indicate the number of structures in that category that have at least one complete biological assembly match according to VAST+.

Currently, ~53% of polypeptide-containing structures in MMDB have >1 polypeptide chain. The histogram plotted in Figure 1 breaks down the numbers by oligomer size and indicates that large fractions of the oligomeric assemblies have, in general, structure neighbors that match the entire assemblies. It should be noted that the fractions might be somewhat exaggerated, as exact duplicates of a structure would be counted as biological assembly matches, and no attempt was made to remove redundant structures or classify biological assembly matches as informative versus uninformative.

THE VAST+ WEB SERVICE

Structure neighbors as computed by the VAST+ algorithm will be used in the future to provide links to 'similar structures' on Entrez/structure document summaries. Lists of similar structures are then summarized via a new interactive web service, which can also be used independently of the Entrez query and retrieval system and provides tools for sub-setting results, at http://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi. For a query structure specified by the user, the service lists similar structures, should they exist, ranked by the extent of the match. Matches that associate each polymer chain of the query with a corresponding polymer chain of some other structure are considered complete and are indicated with full circles in the search results table; partial matches are

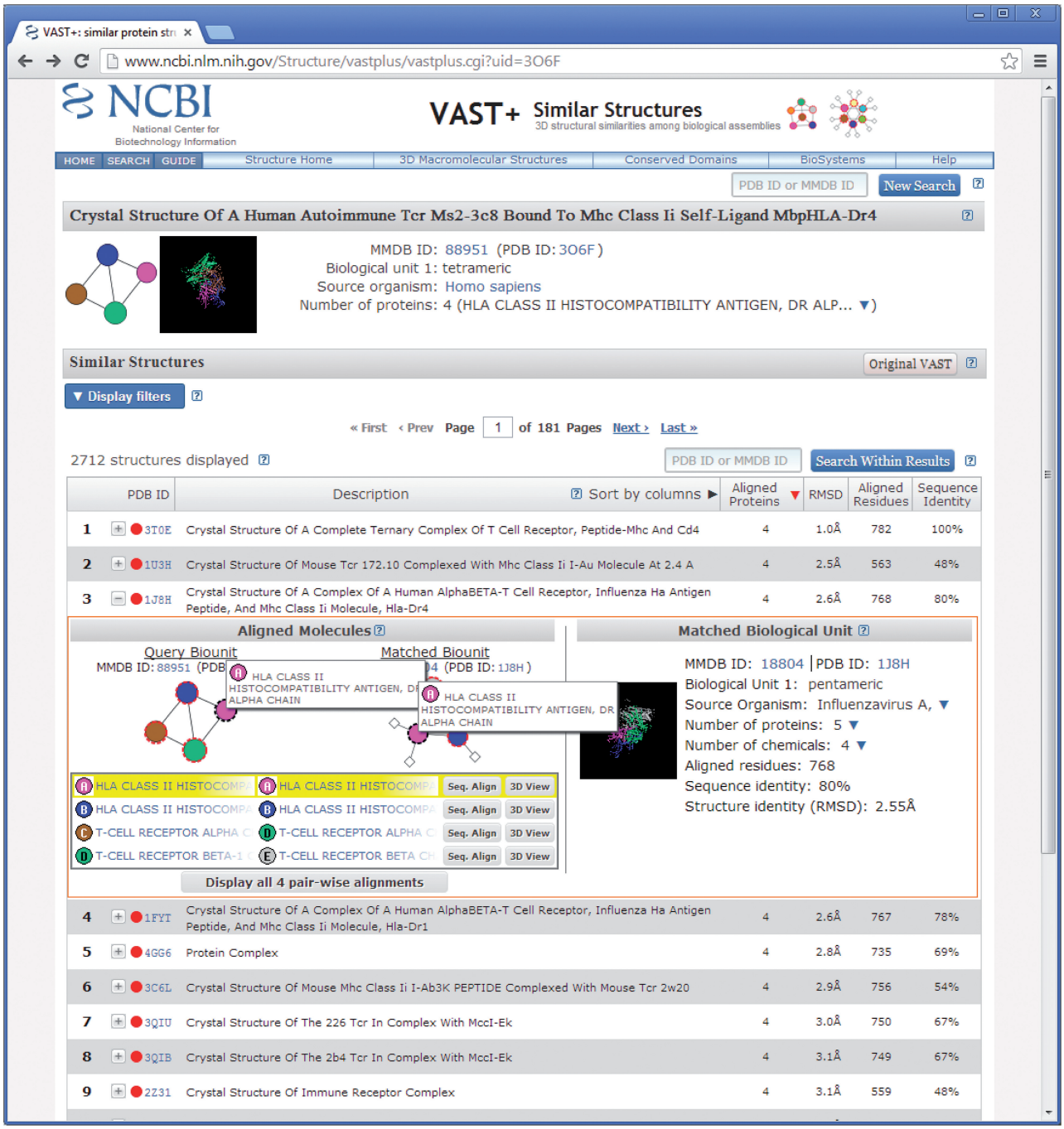

Figure 2. The VAST+ web service generates lists of structures that have 3D similarity to the query. Matches are evaluated with biological assemblies as the unit of comparison (referred to as Biological Units) and may summarize simultaneous alignment of several protein molecule pairs. The query structure '3O6F' (24) currently yields 2712 structure neighbors. Only the 115 neighbors with a complete biological assembly match have been selected in this example (via the 'display filters' menu, shown as collapsed in this figure). The 115 complete matches have been sorted by RMSD, and the third ranking match has been selected to provide more detail. The tabulated matches are shown with their PDB accession, descriptive text, the number of proteins aligned in the match, the total number of aligned residues, the sequence identity and the RMSD resulting from the simultaneous superimposition of all aligned molecules. In this example, the query '3O6F' matches the structure '1J8H' with a total of four aligned protein molecules, totaling 768 residues and resulting in a superposition with 2.55 Å RMSD. 80% of the residues in 3O6F and 1J8H that were spatially aligned by VAST are identical. The extended panel characterizing this selected match contains a table that lists pairs of matching/aligned proteins, and it provides schematic depictions of each biological assembly's composition and interactions. The user can mouse-over those schematics to identify individual molecules and their corresponding match in the other structure (as shown in this example). The individual protein match table contains action buttons that provide access to the pairwise sequence alignments as derived from the VAST superimposition and launch points for visualization of the structure superimposition with the protein structure viewer Cn3D (23). Each "3D View" button will open a superposition of the complete biological assembly alignment with the 3D view centered on the selected protein molecule and its sequence data featured in the Cn3D sequence viewer window. Next to the Aligned Molecules table, an information box lists some stats that characterize the matched biological assembly.

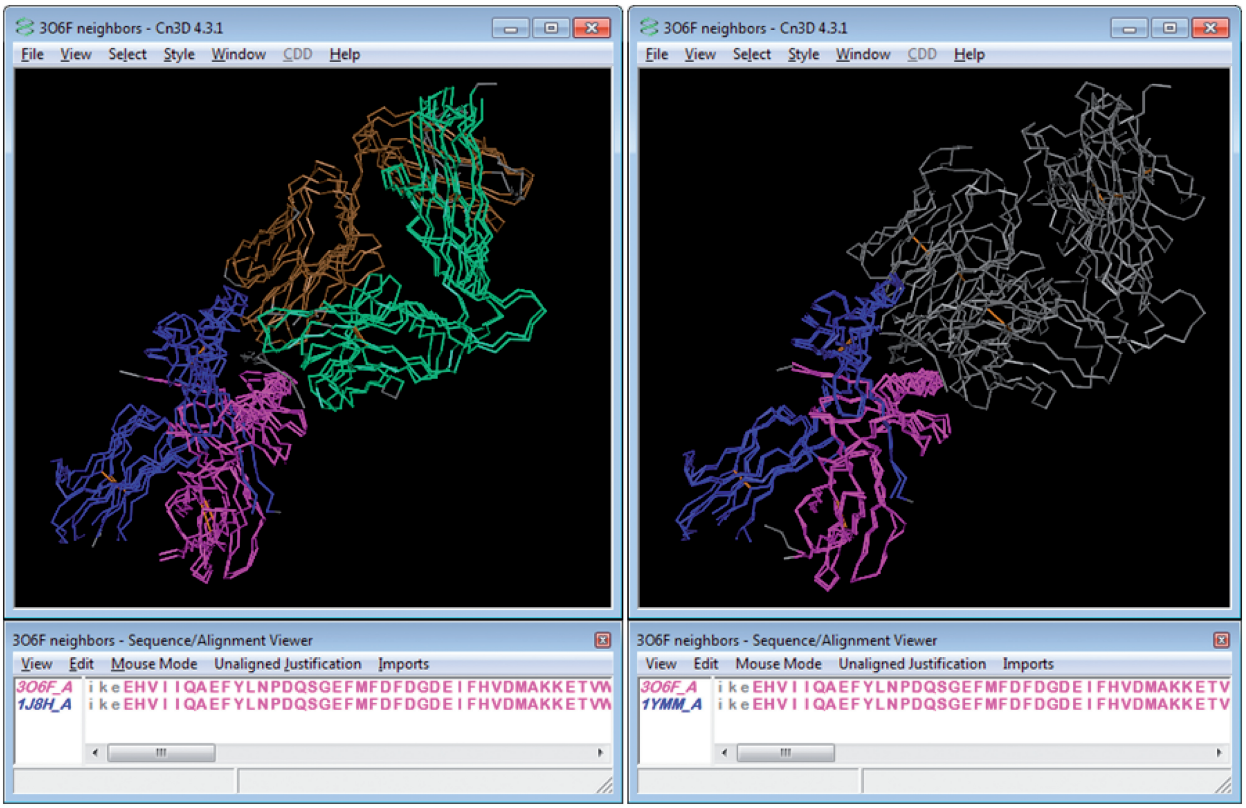

Figure 3. Visualization of structurally matching biological assemblies, as rendered by the visualization tool Cn3D. Cn3D is a helper application for the web browser, available for Windows and OS-X platforms. The query structure, PDB accession 3O6F, represents the complex of an autoreactive T-cell receptor (MS2-3C8, molecules rendered in green and brown) complexed with a self-peptide derived from myelin basic protein and the multiple sclerosis-associated MHC molecule HLA-DR4 (molecules rendered in magenta and blue) (24). The self-peptide has been fused with the MHC molecule for the experiment, which explains why the query is represented as a biological assembly with only four components (Figure 2), and is rendered in gray, as is the default for all unaligned segments in Cn3D visualization sessions launched from VAST+ results pages. The left panel shows 3O6F superimposed with the structure neighbor, PDB accession 1J8H (25), which contains a complex between HLA-DR3, an Influenza hemagglutinin peptide, and a human alpha/beta T-cell receptor. Molecules are rendered so that their colors match those of the corresponding query molecules. The structures of the two complexes match well, resulting in a superimposition of 768 amino acid residues at ~2.6 Å RMSD. This demonstrates how well the autoreactive T-cell receptor complex mimics complexes that include foreign peptides, and it is thought that this binding mode is responsible for the autoimmune TCR escaping negative selection. The right panel shows the VAST+ alignment between 3O6F and the structure of a T-cell receptor from a patient with multiple sclerosis, complexed with a myelin basic protein-derived peptide and an HLA-DR2 MHC, PDB accession 1YMM (26). The conformations of the two complexes are different although their components are similar, and VAST+ does not consider the complete biological assemblies to match. Instead, it reports the most extensive sub-structure match, which in this case involves both subunits of the MHC (molecules rendered in magenta and blue). The molecules corresponding to the TCR are rendered in gray color and would not be displayed by default. The unusual conformation of the complex reported in 1YMM is thought to represent an alternative binding mode that helps autoimmune TCRs to escape negative selection.

indicated with partially filled circles. The default ranking puts matches with the most matched components at the top of the list. Not all queries that have similar structures according to VAST+ are guaranteed to also have complete matches (although monomers usually do). The search results tables provided by the VAST+ web service give a concise summary of the matches and the extent/quality of the similarity. A clickable '+' symbol opens a panel for a selected match that provides more details and functionality.

USING CN3D TO VISUALIZE BIOLOGICAL ASSEMBLY ALIGNMENTS

The 3D structures of superimposed biological assemblies may be visualized using the 3D viewer Cn3D (23), which has been re-released as a new version 4.3.1 to support the visualization style. Currently, Cn3D is able to display the structure superposition of the matched biological assemblies and all the protein chains involved, but it can only display one sequence alignment at a time. Therefore, the individual protein match table as shown in Figure 2 provides separate Cn3D launch points for each matched/aligned protein pair. All of these launch points will result in the same 3D image and rendering, but they will differ in the pair of aligned sequences that are chosen as the content of Cn3D's sequence/alignment viewer window. Figure 3 provides examples of Cn3D visualization sessions. Pairs of matching molecules are rendered in the same color, with unaligned segments rendered in gray. The default rendering settings, as generated and provided by the VAST+ service, can be examined and modified via Cn3D's Style|Annotate menu.

Table 1. URLs for MMDB and VAST resources

MMDB	Database home page	http://www.ncbi.nlm.nih.gov/structure
MMDB FTP	Data distribution	ftp://ftp.ncbi.nih.gov/mmdb/
VAST	Identify structurally similar individual protein molecules	http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
VAST+	Identify structurally similar macromolecular complexes	http://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi
VAST search	Input the 3D coordinates of a query structure to search for similar structures	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html
Cn3D	Molecular graphics viewer	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
CBLAST	Find 3D structures that are related to a query protein via sequence comparison	http://www.ncbi.nlm.nih.gov/Structure/cblast/cblast.cgi

SIMILAR SUBSTRUCTURES: ORIGINAL VAST AND VAST-SEARCH

MMDB is updated weekly, following PDB's schedule. With each update, computation of new structure neighbors is completed within a few days, and they are available as structure neighbors computed for biological assemblies via the VAST+ service, as well as structure neighbors computed for individual protein chains and domains, via the original VAST service. The latter is accessible on the VAST+ pages via a button labeled 'original VAST' that can be found near the top of the VAST+ results page. At this point, the VAST-search service, which accepts 3D structure data uploaded in PDB-format, remains unchanged, and presents similar structures in the Original VAST format (Table 1).

LIMITATIONS AND FUTURE WORK

The current implementation of VAST+, the associated database of pre-computed structure comparison results and the web service represent a first attempt at providing a comprehensive set of structure neighboring information for biological assemblies. Several issues may limit the potential applications of the idea, and we intend to address them in future releases of the service and the associated data. Currently, VAST+ neighboring is restricted to the default biological assembly for each structure, as determined by the content of the structure data and MMDB parsing. Although multiple biological assemblies, if present in a structure entry, tend to be closely similar copies of each other, there are exceptions that should be considered explicitly. Also, VAST+ currently draws on results of VAST neighboring as computed for complete protein molecules and ignores a larger set of results obtained for individual domains. This was done intentionally, so as to speed up the computation, and as the first implementation was intended to focus on structural similarities that are both strong and global. We anticipated that most cases, where two entire biological assemblies can be superimposed globally, would break down into individual protein pairs that can also be aligned and superimposed globally, and not just at the level of individual domains. More importantly, VAST+ makes no attempt at this point at refining the alignment and superpositions after detecting a match between two biological assemblies. It is conceivable that such refinement would, in many cases, results in somewhat shorter alignments and lower RMSD values and might be useful in emphasizing the conserved contact interface between the components of a molecular complex. Furthermore, a strategy for detecting biological assembly matches that considers multiple molecules simultaneously might exhibit higher sensitivity and pick up similarities that cannot be found via the 2-tiered approach we have presented here. Currently, VAST+ ignores non-polypeptide components of macromolecular complexes, but certainly both nucleic acids and chemical ligands, if present, could be matched as well as the protein components. It should be mentioned that no VAST+ neighboring data are available for structure database entries that lack assignment of biological assemblies, and currently VAST+ skips biological assemblies whose size exceeds a threshold number of protein components—the systematic evaluation of all possible multimolecule matches becomes too time-consuming, and will need to be supplemented by a suitable heuristic.

FUNDING

Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS. Comments, suggestions and questions are welcome and should be directed to: info@ncbi.nlm.nih.gov. Funding for open access charge: Intramural Research Program of the National Library of Medicine at the National Institutes of Health/DHHS.

Conflict of interest statement. None declared.

REFERENCES

1. Madej,T., Addess,K.J., Fong,J.H., Geer,L.Y., Geer,R.C., Lanczycki,C.J., Liu,C., Lu,S., Marchler-Bauer,A., Panchenko,A.R. *et al.* (2012) MMDB: 3D structures and macromolecular interactions. *Nucleic Acids Res.*, **40**, D461–D464.
2. Rose,P.W., Bi,C., Bluhm,W.F., Christie,C.H., Dimitropoulos,D., Dutta,S., Green,R.K., Goodsell,D.S., Prlic,A., Quesada,M. *et al.* (2013) The RCSB protein data bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.
3. Gibney,G. and Baxevanis,A.D. (2011) Searching NCBI databases using Entrez. *Curr. Protoc. Hum. Genet.*, **Chapter 6**, Unit 6.10.
4. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
5. Marchler-Bauer,A., Zheng,C., Chitsaz,F., Derbyshire,M.K., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M., Hurwitz,D.I., Lanczycki,C.J. *et al.* (2013) CDD: conserved domains and

protein three-dimensional structure. *Nucleic Acids Res*., **41**, D348–D352.
6. Shoemaker,B.A., Zhang,D., Tyagi,M., Thangudu,R.R., Fong,J.H., Marchler-Bauer,A., Bryant,S.H., Madej,T. and Panchenko,A.R. (2012) IBIS (Inferred Biomolecular Interaction Server) reports, predicts, and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res*., **40**, D834–D840.
7. Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol*., **372**, 774–797.
8. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol*., **6**, 377–385.
9. Haseqawa,H. and Holm,L. (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol*., **19**, 341–348.
10. Prlic,A., Bliven,S., Rose,P.W., Bluhm,W.F., Bizon,C., Godzik,A. and Bourne,P.E. (2010) Pre-calculated protein structure alignments at the RCSB PDB website. *Bioinformatics*, **26**, 2983–2985.
11. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classficiation of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.
12. Sillitoe,I., Cuff,A.L., Dessailly,B.H., Dawson,N.L., Furnham,N., Lee,D., Lees,J.G., Lewis,T.E., Studer,R.A., Rentzsch,R. *et al*. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res*., **41**, D490–D498.
13. Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19(Suppl. 2)**, ii246–ii255.
14. Holm,L. and Rosenstrom,P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res*., **38(Suppl. 2)**, W545–W549.
15. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol*., **313**, 903–919.
16. Orengo,C.A. and Taylor,W.R. (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol*., **266**, 617–635.
17. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*., **11**, 739–747.
18. Marchler-Bauer,A. and Bryant,S.H. (1997) Measures of threading specificity and accuracy. *Proteins*, **(Suppl. 1)**, 74–82.
19. Sierk,M.L. and Pearson,W.R. (2004) Sensitivity and selectivity in protein structure comparison. *Protein Sci*., **13**, 773–785.
20. Kim,C. and Lee,B. (2007) Accuracy of structure-based sequence alignments of automatic methods. *BMC Bioinformatics*, **8**, 355.
21. Sippl,M.J. and Wiederstein,M. (2012) Detection of spatial correlations in protein structures and molecular complexes. *Structure*, **20**, 718–728.
22. Mukherjee,S. and Zhang,Y. (2009) MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res*., **37**, e83.
23. Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci*., **25**, 300–302.
24. Yin,Y., Li,Y., Kerzic,M., Martin,R. and Mariuzza,R.A. (2011) Structure of a TCR with high affinity for self-antigen reveals basis for escape from negative selection. *EMBO J*., **30**, 1137–1148.
25. Hennecke,J. and Wiley,D.C. (2002) Structure of a complex of the human alpha/beta T cell receptor (TCR) HA1.7, Influenza hemagglutinin peptide, and major histocompatibility complex class II molecule, HLA-DR4 (DRA*0101 and DRB1*0401): insight into TCR cross-restriction and alloreactivity. *J. Exp. Med*., **195**, 571–581.
26. Hahn,M., Nicholson,M.J., Pyrdol,J. and Wucherpfennig,K.W. (2005) Unconventional topology of self-peptide major histocompatibility complex binding by a human autoimmune T cell receptor. *Nat. Immunol*., **6**, 490–496.